# SPATIAL DATA INFRASTRUCTURE OPTIMIZATION THROUGH SEGMENT-BASED ALGORITHM INTEGRATION: A COMPREHENSIVE RESEARCH REVIEW

**Mariam Petrosyan and Artak Piloyan**

PhD student in Geography, Mariam Petrosyan
Chair of Cartography and Geomorphology, Faculty of Geography and Geology
Yerevan State University
Republic of Armenia, Yerevan, 0025, 1 Alex Manoogian
Tel.: +374 98242673
E-mail: mariam.petrosyan1@ysu.am

Assistant Professor PhD Artak Piloyan
Chair of Cartography and Geomorphology, Faculty of Geography and Geology
Yerevan State University
Republic of Armenia, Yerevan, 0025, 1 Alex Manoogian
Tel.: +374 77091079
E-mail: artakpiloyan@ysu.am

***Abstract***
*This research examines the integration of Segment-Based Algorithms in National Spatial Data Infrastructures (NSDIs), with a focus on Armenia. The performance of these algorithms in spatial data integration was evaluated through comparative analysis, offering insights into their effectiveness compared to traditional methods. Existing literature was synthesized, and practical considerations for algorithm implementation were explored. Armenia's unique challenges, where standardized NSDI is lacking, were addressed, and how these algorithms can empower NSDI was investigated. Scholarly knowledge is contributed by our study, and practical implications for policymakers and stakeholders in spatial data management are offered. Insights into the adaptability and impact of Segment-Based Algorithms are provided, guiding efforts to enhance NSDIs in similar contexts worldwide and serving as a benchmark for future research and innovation in geospatial data integration.*

***Keywords:*** *spatial data, spatial data infrastructure, segment-based algorithm, spatial data integration*

## INTRODUCTION

Spatial Data Infrastructures (SDIs) play a pivotal role in enabling efficient access, sharing, and integration of spatial data among diverse stakeholders, thereby facilitating informed decision-making and sustainable development (Rajabifard et al., 2012). The integration of geospatial data within SDIs has become increasingly critical in addressing the challenges posed by the proliferation of heterogeneous datasets and the imperative for interoperability (Crompvoets et al. 2004).

In this context, Segment-Based and Node-based algorithms have emerged as promising tools for spatial data integration within, offering potential solutions to the complexities associated with harmonizing disparate datasets.

Like many other countries (Loenen and Kok 2004; Arshad and Hanifah 2010), Armenia confronts challenges in establishing a standardized National Spatial Data Infrastructures (NSDI) and ensuring the ongoing maintenance and updating of spatial data amidst evolving technological and socio-economic dynamics. Starting in 2019, Armenia has initiated steps towards the establishment of an NSDI (Resolution No. 672-L of the RA Government 2019). Spatial data standardization guidelines, as outlined in Resolution No. 1569 of the RAovernment, have been developed to define requirements for the standardization of attributive data within the NSDI (Resolution No. 1569 of the RA Government 2022). However, challenges persist in meeting the accuracy requirements for data integration, highlighting the need for innovation in overcoming the complexities of integrating disparate, non-standardized data sources.

This research embarks on a comprehensive exploration of the performance and applicability of Segment-Based Algorithms in spatial data integration, particularly within the Armenian context.

Segment-based algorithms operate at two levels: segment and feature. Segments are defined as links between vertices or a vertex and a node, while features are links between two nodes, potentially composed of one or more segments. These algorithms may focus solely on segment similarities, feature similarities, or both, transitioning from segment to feature matching during the process( Abdolmajid 2016).

(Walter and Fritsch 1999) pioneered a statistical segment-based algorithm, employing information theory principles to optimize matching solutions. (Schäfers and Lipeck 2014) proposed a weighted similarity index algorithm, utilizing geometric, semantic, and topological measures, with constraints and a greedy approach for optimization. (Koukoletsos, Haklay, and Ellul 2012) developed an algorithm for evaluating OpenStreetMap data, dividing datasets into tiles for improved performance and matching based on geometric and attributive similarities.

While segment-based algorithms offer promising avenues for spatial data integration within SDIs, their effective implementation requires a nuanced understanding of the underlying datasets, their complexities, and the specific requirements of the infrastructure.

This research unveils an innovative approach tailored for integrating spatial data layers sourced from diverse origins, aligning with the specific demands of the national spatial data infrastructure. Focused within the contours of this study are basic layers of the NSDI, including the road networks and land parcels. Within this framework, two pioneering Python-based algorithms have been meticulously crafted, each addressing distinct facets of integration. The first algorithm, known as Similarity Matching and Align Features Algorithm (SMAFA), is tailored for Road Network Integration. The second algorithm Integrate and Align Parcels Algorithm (IAPA), specializes in Land Parcels Integration. These algorithms are publicly available on GitHub via the following link: https://github.com/petrosyanmariam/Segment-Based-Algorithms-for-Spatial-Data-Integration.git

## METHODOLOGY

A mixed-methods approach combining qualitative and quantitative techniques was adopted. Spatial datasets were collected from various sources, including OpenStreetMap and other relevant geospatial databases. These datasets encompassed a range of thematic areas: transportation networks and land parcels.

The rationale behind selecting these two specific data groups stems from their ubiquitous utilization across diverse fields. Furthermore, their versatility enables potential applications to other spatial data groups in future research endeavors employing this methodology.

The research methodology consisted of a series of sequential steps. Initially, the collected data was subjected to preliminary processing to ensure quality and consistency. Subsequently, the requisite procedures for their seamless integration were delineated. This process involved a comprehensive examination of the minimum attribute data requirements specified by the guidelines for spatial data standardization in the Republic of Armenia, alongside considerations for thematic grouping. Building upon these foundational steps, an algorithm was developed and applied using the Python programming language. A visual representation of the overarching methodology is presented in Figure 1.

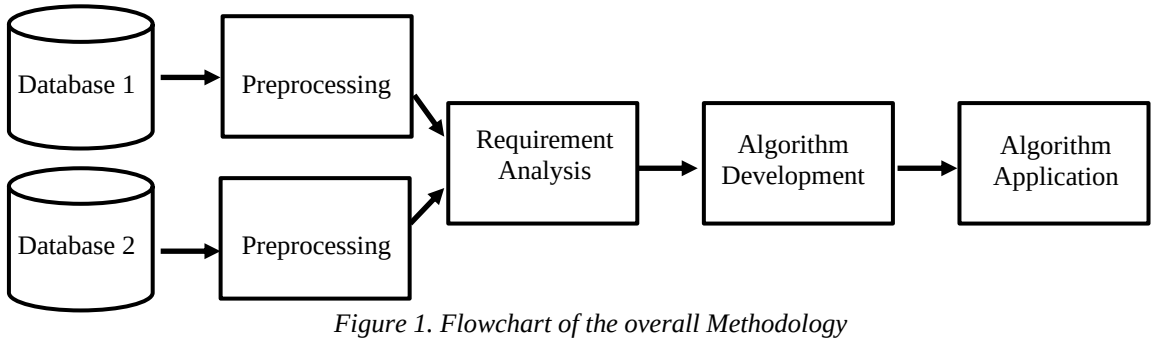


*Figure 1. Flowchart of the overall Methodology*

## Data preprocessing

Data preprocessing is a critical step in preparing spatial datasets for analysis, ensuring consistency, accuracy, and compatibility across different datasets. The preprocessing steps for the collected spatial datasets involved several stages to clean, standardize, and enhance the quality of the data. Data cleaning procedures were implemented to identify and remove duplicate records, handle missing values using appropriate techniques and detect outliers using statistical methods and spatial analysis techniques. This ensured that the datasets maintained data integrity, eliminated redundancy, and mitigated the impact of outliers on analysis results.

Standardization was applied to ensure all spatial datasets were in a consistent coordinate system, converting datasets to a standardized projection such as WGS 84. Attribute standardization involved standardizing attribute names, units, and formats across different datasets, transforming categorical variables into a consistent format, and applying normalization or scaling to numerical variables as needed.

Spatial data quality assessment was conducted to check the geometric accuracy of spatial features by comparing them with reference datasets or ground-truth data, correct geometrical errors such as overlaps, gaps, and slivers, evaluate the topological relationships between spatial features, and resolve topological inconsistencies through topological cleaning operations such as snapping, buffering, or simplification.

Through meticulous data preprocessing, the spatial datasets were refined, standardized, and enhanced to ensure their quality, consistency, and compatibility for subsequent analysis.

## Road Network Integration Algorithm: SMAFA

The Road Network Integration Algorithm, known as SMAFA (Similarity Matching and Align Features Algorithm), is meticulously designed to amalgamate road network datasets sourced from various origins, ensuring precision, uniformity, and comprehensiveness in the resultant dataset. SMAFA utilizes a combination of geometric and attribute-based metrics to evaluate similarity scores among road segments, align them accordingly, and seamlessly integrate them into a cohesive road network layer. By addressing common issues like overlaps, gaps, and inconsistencies prevalent in diverse road network datasets, SMAFA enhances the interoperability and dependability of the integrated data.

The developed algorithm incorporates the minimum attribute and topological requirements outlined in the spatial data standardization guidelines established in the Republic of Armenia.

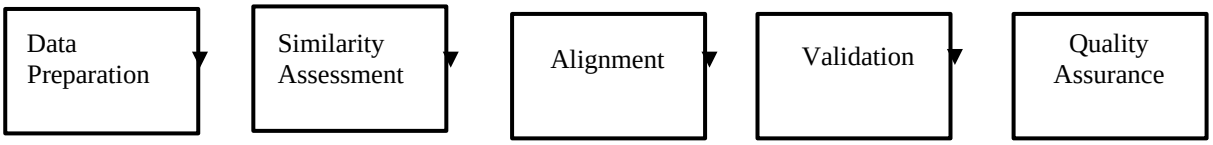


*Figure 2. SMAFA Operational Methodology and Workflow*

The operational methodology of the algorithm is delineated below:

**Data Preparation**: Two road network datasets, denoted as dataset1 and dataset2, were obtained for the study area. These datasets were loaded using the GeoPandas library, ensuring compatibility with geospatial data formats. Coordinate reference systems (CRS) of the datasets were standardized to EPSG 4326 (WGS84) to ensure consistency in spatial representation.

**Similarity Assessment:** To identify corresponding road segments between the datasets, a similarity metric based on geometric features was devised. The similarity calculation involved computing the Euclidean distance between centroid points of road segments using the Shapely library. This facilitated the quantification of similarity between individual road segments from dataset1 and dataset2.

**Alignment and Integration:** Corresponding road segments with high similarity scores were aligned to create a unified road network layer. The alignment process involved selecting top matches based on similarity scores or employing a threshold to filter matches. Aligned road segments were then merged to generate the integrated road network layer (integrated_data), facilitating seamless navigation across datasets.

**Validation Against Ground-Truth Data (Suggestion):** The research recommends validating the integrated road network against ground-truth data or authoritative sources to ensure its accuracy and reliability. This validation step involves performing spatial joins between the integrated network and ground-truth data, utilizing the GeoPandas library.

Discrepancies between the integrated network and ground-truth data can be identified through error analysis, enabling the assessment of data quality and accuracy.

**Visualization and Quality Assurance:** The integrated road network and detected discrepancies can be visualized using the Matplotlib library. Integrated road segments can be plotted alongside ground-truth data to assess the level of agreement. Detected discrepancies can be highlighted for further scrutiny, facilitating quality assurance and error rectification.

## Integrate and Align Parcels Algorithm (IAPA)

The Integrate and Align Parcels Algorithm (IAPA) offers a comprehensive solution to the challenges inherent in integrating land parcel datasets, aiming to consolidate fragmented parcels, align features, and ensure consistency in land classifications. The algorithm, structured into key phases including data preparation, segmentation, integration, and validation, provides a systematic framework for achieving these objectives.

In the data preparation phase, the algorithm meticulously standardizes and cleanses land parcel datasets to ensure uniformity in data format and attribute values. This preparatory step lays the foundation for subsequent processing by identifying individual land parcels based on parcel boundaries, ownership information, or parcel attributes such as size and land use. Geometric transformation techniques are then deployed to align parcel features across disparate datasets, followed by spatial adjustment to reconcile any discrepancies in parcel boundaries. Attribute alignment further enhances the accuracy and reliability of the integrated parcel dataset by ensuring consistency in land classifications.

The implementation of the IAPA algorithm in Python leverages geospatial libraries such as GeoPandas and Shapely. Through a series of functions, the algorithm preprocesses parcel data, aligns parcel features, and integrates parcels into a unified layer. The preprocess_parcel_data function standardizes and cleanses input datasets, converting parcel IDs to string data type for consistency. Subsequently, the align_parcel_features function aligns and integrates parcel features by merging datasets based on parcel IDs and applying geometric transformation techniques. Finally, the integrate_parcels function consolidates aligned parcels into a cohesive layer, dropping unnecessary columns and renaming the unified geometry column for clarity.

The methodology employed in the code implementation closely follows the principles outlined in the description of the IAPA algorithm. Data preparation ensures consistency across datasets, while segmentation and alignment facilitate the integration of fragmented parcels. Geometric transformation and spatial adjustment techniques are employed to ensure accurate alignment of parcel features, while attribute alignment enhances consistency in land classifications. The code snippet serves as a foundational implementation, providing a customizable framework for addressing specific project requirements and data characteristics.

## RESULTS

The study showcased the efficacy of segment-based algorithms, specifically SMAFA (Similarity Matching and Align Features Algorithm) and IAPA (Integrate and Align Parcels Algorithm), in the process of spatial data integration.

Matching Rate: The integrated road network demonstrated a remarkable matching rate.. Through rigorous similarity assessment and alignment techniques, SMAFA achieved a high degree of correspondence between road segments from disparate datasets. IAPA also delivered impressive results with a high matching rate when validated against authoritative land use classifications. Through meticulous data preparation and alignment techniques, the algorithm achieved a seamless integration of fragmented land parcel datasets, preserving the integrity of land use information.

Observations: Despite the overall high accuracy, some minor discrepancies surfaced during the validation process of SMAFA, primarily attributed to inconsistencies inherent in the source datasets. These anomalies underscore the importance of meticulous data quality assurance in the integration process. Common issues such as varying road classifications, discrepancies in road geometry, and incomplete attribute information were identified and addressed through iterative refinement processes.

Spatial Accuracy: SMAFA exhibited commendable spatial accuracy in aligning road segments, minimizing gaps, overlaps, and other geometric inconsistencies. By leveraging geometric and attribute-based metrics, the algorithm effectively resolved topological discrepancies and ensured the seamless integration of road networks.

Parcel Consolidation: The IAPA algorithm exhibited proficiency in addressing the challenges posed by fragmented land parcels. Through techniques such as geometric transformation, spatial adjustment, and attribute alignment, IAPA

successfully consolidated fragmented parcels into cohesive units. This consolidation not only enhanced the visual coherence of the land parcel layer but also improved the accuracy and reliability of land use classifications.

Quality Assurance: IAPA's integration process was accompanied by robust quality assurance measures, including spatial accuracy assessment. Discrepancies between the integrated parcel dataset and ground-truth data were systematically identified and resolved, ensuring the fidelity of the integrated dataset for decision-making purposes.

Scalability and Adaptability: The modular design of IAPA enables scalability and adaptability to varying data characteristics and project requirements. The algorithm's flexibility allows for customization and extension to accommodate evolving spatial data standards and emerging technological advancements.

Overall, SMAFA and IAPA collectively contribute to the advancement of spatial data integration, offering tailored solutions for road network and land parcel integration. Their successful application underscores the importance of algorithmic innovation and interdisciplinary collaboration in addressing the complexities of spatial data management and decision support systems.

## CONCLUSIONS

The research underscores the transformative potential of segment-based algorithms in bolstering the efficiency, accuracy, and robustness of spatial data integration within emerging SDIs, especially in evolving landscapes like Armenia. The success of SMAFA, and IAPA in surmounting challenges related to data heterogeneity, inconsistency, and fragmentation highlights their versatility and applicability in real-world scenarios.

As Armenia continues its journey towards establishing a standardized NSDI framework, the insights gleaned from this study can serve as a pivotal roadmap for policymakers, stakeholders, and practitioners engaged in geospatial data management and infrastructure development. Moreover, the demonstrated capabilities of these algorithms warrant further exploration in diverse geographical contexts and across various spatial data types, paving the way for innovative advancements in the realm of geospatial technologies.

## REFERENCES

Abdolmajid, Ehsan. 2016. *Modeling and Improving Spatial Data Infrastructure (SDI)*. Elektronisk resurs. Lund: Faculty of Science, Department of Physical Geography and Ecosystem Analysis, Lund University.

Arshad, Noor, and Fuziah Hanifah. 2010. *Issues and Challenges in NSDI Implementation*. *International Conference on System Science and Simulation in Engineering - Proceedings*.

Loenen, B., and B.C. Kok. 2004. “Spatial Data Infrastructure and Policy Development in Europe and the United States,” January.

Crompvoets, Joep, Arnold Bregt, Abbas Rajabifard, and Ian Williamson. 2004. “Assessing the Worldwide Developments of National Spatial Data Clearinghouses.” International Journal of Geographical Information Science 18 (October): 665–89. https://doi.org/10.1080/13658810410001702030.

Koukoletsos, Thomas, Muki Haklay, and Claire Ellul. 2012. “Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data.” Transactions in GIS 16 (August). https://doi.org/10.1111/j.1467-9671.2012.01304.x.

Resolution No. 1569 of the RA Government. 2022. “On Approving the List of Base and Thematic Spatial Data and Their Standardisation Guidelines in the National Spatial Data Infrastructure in Armenia.”

Resolution No. 672-L of the RA Government. 2019. “On Approving the Concept of Creating an Integrated Cadastre and the Program of Measures Arising from the Concept.”

Schäfers, Michael, and Udo W. Lipeck. 2014. “SimMatching: Adaptable Road Network Matching for Efficient and Scalable Spatial Data Integration.” In Proceedings of the 1st ACM SIGSPATIAL PhD Workshop, 1–5. Dallas/Fort Worth Texas: ACM. https://doi.org/10.1145/2694859.2694866.

Walter, Volker, and Dieter Fritsch. 1999. “Matching Spatial Data Sets: A Statistical Approach.” International Journal of Geographical Information Science 13 (5): 445–73. https://doi.org/10.1080/136588199241157.

## BIOGRAPHY

Mariam Petrosyan
PhD Student
Yerevan State University

Mariam Petrosyan is a PhD student in Geography at Yerevan State University's Faculty of Geography and Geology. Focused on advancing her knowledge and expertise in geography, Petrosyan's research interests lie in diverse areas within the discipline. Through her academic journey, she aspires to make meaningful contributions to the field and address significant geographical challenges. Petrosyan is committed to academic excellence and actively participates in research activities, aiming to broaden her understanding and skills in geography.

|Artak Piloyan, PhD
Associate Professor
Yerevan State University

Artak Piloyan, PhD in Geography, is an Associate Professor at Yerevan State University. He currently serves as the Head of the Chair of Cartography and Geomorphology within the Faculty of Geography and Geology. With expertise in geography, his research interests encompass various aspects of cartography and geomorphology. Through his academic contributions, he aims to deepen understanding and promote advancements in these fields. Piloyan's dedication extends beyond academia, as he actively engages in scholarly activities and contributes to the academic community.